REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From – To)*
06-11-2003	Final Report	30 September 2002 - 30-Sep-03

4. TITLE AND SUBTITLE

Adding Scale to the Modulation-Frequency Transform

5a. CONTRACT NUMBER
FA8655-02-1-3083

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Dr. Roy Dunbar Patterson

5d. PROJECT NUMBER

5d. TASK NUMBER

5e. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Cambridge
Downing Street
Cambridge CB2 3EG
United Kingdom

8. PERFORMING ORGANIZATION REPORT NUMBER
N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

EOARD
PSC 802 BOX 14
FPO 09499-0014

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
SPC 02-4083

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This report results from a contract tasking University of Cambridge as follows: The contractor will investigate and compare the first two stages of the AFMFT ("Acoustic-Frequency", "Modulation-Frequency" Transform) and SWMT (Stabilized Wavelet Modulation-frequency Transform) and to produce an optimal, three-stage frequency-scale transform for use in coding and speech recognition. This will be used for (1) Automatic classification of natural and mechanical sounds that vary in scale and (2) Automatic speech recognition in noisy and reverberant environments. The primary deliverables will be (a) a report describing the development of a scale database for speech sounds with a CD of the database, (b) a report describing a psychophysical study of human performance with respect to scale in vowel sounds. If time and resources allow, additional deliverables may include (c) a report describing the development of a scale database for musical instrument sounds with a CD of the database, (d) a report describing a psychophysical study of human performance with respect to scale in musical sounds.

15. SUBJECT TERMS
EOARD, Signal Processing, Speech Processing

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18, NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			PAUL LOSIEWICZ, Ph. D.
UNCLAS	UNCLAS	UNCLAS	UL		19b. TELEPHONE NUMBER *(Include area code)* +44 20 7514 4474

The perception of scale in vowels

David R. R. Smith[a] and Roy D. Patterson[b]

Centre for the Neural Basis of Hearing,
Department of Physiology,
University of Cambridge,
Downing Street, Cambridge CB2 3EG
United Kingdom

[a] david.smith@mrc-cbu.cam.ac.uk, [b] roy.patterson@mrc-cbu.cam.ac.uk

SUMMARY Previous reports presented our psychophysical findings on the perception of vowels which had been manipulated to make them sound like smaller and larger people, including some well beyond the normal range of the population. This final report includes, in addition to this previous research, an experiment showing speaker size can be extracted from a speech-like sequence of vowels that does not possess any simple spectral cue. We provide a detailed motivation and discussion of scale in vowel sounds. Our results show that we can be confident that human listeners are able to extract both vowel type and speaker size from vowel sounds even when the size and pitch are well beyond normal experience.

ABSTRACT

The resonating properties of many objects provide acoustical correlates which can be used to gain information about the objects. The acoustic signal provides not only shape information (what the sound means) but also size information (how small/big the object is relative to the population). A signal processing algorithm able to isolate both shape and size information is the Mellin transform. It is posited that such a transform is applied to all sounds at an early point in the auditory system (Irino and Patterson, 2002). Our ability to tell what vowel was spoken (vowel normalisation), despite gross waveform changes due to different vocal tract lengths and larynx size differences across sex and age, shows we are able to extract the invariant vowel qualities (shape information). We are also able to tell whether the speaker was a man, woman or child (size information). We manipulated vowel sounds to have the acoustical properties of different size speakers scaled way beyond the usual range of variation in the population. Listeners were able to both identify the vowels and extract information about the size of the speaker of these scaled vowels. Our results are unlikely to be due to some speech-specific learnt statistical correlation between pitch and formant frequency as the learnt association should fail outside the training set region. The huge range over which listeners were able to extract shape and size information suggests rather the operation of some active pre-processing transform that is applied to all input sounds.

1. INTRODUCTION

Striking an object sets up a series of rebounding (resonating) pressure waves within the internal cavities of the object. The size and shape of the cavities cause certain frequencies to be reinforced or attenuated – imparting a characteristic acoustic signature to the resulting sound wave. The resonant properties of the object thus convey valuable acoustical information which can be potentially used to characterise the object. The sound contains information about both the shape of the object and the relative size of the object (how big/small it is relative to the population). Given the potential advantage of being able to both read and send size information (either true or false estimates of), it would not be surprising if there was selective evolutionary pressure for the brain to develop ways of extracting such information.

Shape and size information is available in the acoustic signature of both inanimate and animate objects. One only has to strike a large beer-barrel and then a small tin-can to appreciate the different sounds they produce, and how that sound conveys information about their relative size. In the animal world there are plentiful examples of size information advertising. Female Fowler toads select mating partners on the basis of the vocal calls of the male toads, with the females preferring larger males (Fairchild, 1981). 'Size exaggeration' can be achieved by dropping the larynx thereby increasing the length of the supra-laryngeal vocal tract (Ohala, 1984). This might account for the marked descent of the larynx in human males (Fitch and Giedd, 1999). Examples from other species include birds coiling their trachea in their body thereby greatly lengthening their vocal tract (Fitch, 1999); red and fallow male deer dropping their larynx deep down into their thoraxes while roaring (Fitch and Reby, 2001), and possible anatomical specialisations in roaring versus non-roaring big cats (Hast, 1989). The basis for truth in size advertising is supported by the positive correlation between vocal tract length and body size in dogs, monkeys and humans (Riede and Fitch, 1999; Fitch, 1997; Fitch and Giedd, 1999). However, it is in the complex world of human vocal communications (speech) that we find a particularly striking example of a domain where size and shape information are used.

A. Human vocal communication

Human listeners can identify specific vowels regardless of whether they are spoken by men, women or children. They also know whether the speaker is a man, woman or child. We are able to do this even though the sounds waves of the specific vowel sounds can be very different for the different groups. For example, if we compare the vowel /a/, as spoken by an adult male and a six-year old female child, we can see that they vary considerably in both the repetition rate (voice pitch) and the frequencies of the most prominent spectral peaks (formants) *cf.* Figure 1. For a given fixed vowel, the differences in voice pitch and formant frequency are largely due to differences in the size of larynx and the vocal tract length (VTL) between men, women and children. Somehow, the auditory system automatically extracts from the sound both the specific vowel spoken (shape information) and whether a man, woman or child spoke it (size information).

What is the anatomical basis for the differences in the vowels of men, women and children? To understand this we need to know how the complex tonal sounds of speech are produced and how the anatomical differences between different sexes and

ages affects this process (*cf*. Figure 2 schematic). Speech occurs when the air stream from the lungs, after being broken up into a series of glottal pulses by opening/shutting action of the vocal folds, excites the VT. Each glottal pulse results in a puff of air that is trapped within the oral cavities. The size and shape of the VT causes certain frequencies to be reinforced and others to be attenuated[1]. The length of the supra-laryngeal VT increases with both sex and age (Fig. 2, 1st column thick black line). The longer the VT, the more the formant frequencies are shifted towards lower frequencies. Compare the adult male formants to the adult female formants to the child's formants (Fig. 2, 2nd column). When plotted on a log frequency axis, the spacing of the formants to each other remains the same – we therefore have a *shifting* rather than a stretching, as would occur on a linear frequency axis. The resulting sounds are of the same vowel but spoken by different speakers. The magnitude spectra (Fig. 2, 3rd column) show how the spectral envelope changes across speakers with the harmonics of the speech sound filling the spectral envelope.

As a child grows between the ages of 4 to 12 (puberty) there is a steady increase in VTL with a correlated decrease in the formant frequencies. There is little sexual dimorphism between these ages. After onset of puberty, males undergo increases in VTL exceeding that predicted by body size alone while female VTL remains correlated to body size. This secondary sexual characteristic means that the formant frequencies of mature males decrease by about 32% from their values at age 4 while the formant frequencies of mature females decrease by about 20% (Fitch and Giedd, 1999; Huber, Stathopoulos, Curione, Ash and Johnson, 1999).

The link between voice pitch and speaker size is less clear. Certainly, there is a strong link between speaker *sex* and pitch. VTL and pitch have about the same value in classifying speaker sex but VTL is much more efficacious than pitch when classifying individual speakers (Bachorowski and Owren, 1999). The sexual dimorphism in pitch is attributable to the puberty-linked increase in testosterone which stimulates growth in the laryngeal cartilages (Beckford, Rood and Schaid, 1985). However, there is no direct correlation between body size and pitch (e.g. Lass and Brown, 1978). This is to be expected because the VTL is directly dictated by the size of the cranium whilst the vocal folds are free-floating of any bony structure (Negus, 1949). We also vary pitch to make prosodic distinctions, e.g. the rising pitch contour of the interrogative sentence, so less linkage should be expected between pitch and body size. Nevertheless, it seems clear that pitch can be used to correctly classify the sex at least of speakers (Bachorowski and Owren, 1999).

B. Learnt statistical correlation or Mellin transform?

Figures 1 and 2 show that the same vowel is acoustically carried by very different sound waveforms. How can the auditory system extract the vowel invariance? There are two basic candidate theories. One theory holds that the auditory system has learnt the statistics of variation in pitch and formant frequency (e.g. Assmann, Nearey and Scott, 2002). The vowel /a/, as spoken by a child and an adult male, are heard as the same vowel because the auditory system has learnt that a high pitch (child) is

[1] The *shape* of the VT is largely determined by the placement of the tongue within the oral cavity. The shape affects the positioning of the formants in frequency relative to each other – different vowels having different vector angles in a multi-dimensional vowel space. For the purposes of our argument we assume the same fixed VT shape across all speakers, i.e. the speakers are uttering the same vowel.

correlated with high formant frequencies. The pattern can be learnt independent of frequency shifts and then used to identify the vowel.

The other theory relies on the reasoning that size is a physical attribute of a sound (in much the same way as its frequency content) and can be recovered with a suitable transform. Kicking a tin-can and rolling a beer barrel down the road produce very different sounds. However, we would easily be able to tell which sound was caused by the small object and which sound was caused by the large object. The Mellin transform (Cohen, 1993) is a signal processing algorithm that is able to segregate both shape and size information. It has been suggested that a form of this transform is applied to all sounds at a relatively early point in the auditory system (Irino and Patterson, 2002) before specific speech recognition processing begins. The Mellin transform maps all input sounds to a nominal scale (allowing access to shape information) whilst encoding size information separately. In the process, the Mellin transform normalises vowels for VTL.

These theories are open to experimental verification. If vowel perception is learnt statistical variation then it should break down if we move beyond the region of normal variation in the human population. An active re-scaling process, used to transform all sounds, might be expected to work across a much wider region of changes in pitch and VTL. Recent work has measured identification performance for vowels manipulated by shifting the range of frequencies or changing the pitch (Fu and Shannon, 1999; Assmann *et al*., 2002). Assmann *et al* used the same vocoder as we did to manipulate their vowels. However, we sample the pitch-VTL space more finely and over a wider range to generate a 2D surface map of vowel identification performance.

We propose to map the region over which people can reliably identify vowels and determine when performance breaks down. To do this we will generate scaled English vowels and ask listeners to identify the vowel spoken. A second aspect of this study will be to measure the discrimination performance (sensitivity) along the pitch and VTL dimensions. Can listeners extract and use the size information in scaled vowels?

II. METHOD

A. Stimuli and equipment

We collected examples of the canonical English vowels (/a/, /e/, /i/, /o/, /u/), as spoken by RP in natural /hVd/ sequences (*hard, heed, hayed, hoed, who'd*) recorded using a high quality microphone (SM58-LCE, Shure). The vowels were sustained (*haaard* etc) to provide long duration vowels. The waveforms were digitised to 'wav' files with 16-bit amplitude resolution and a sampling rate of 44.1 kHz. The vowels were excised out of the /hVd/ sequences, preserving the natural initial onset of the vowel whilst avoiding the preceding /h/ sound. A cosine-squared amplitude function (5 ms onset, 30 ms offset, 565 ms plateau) was used to gate each vowel to avoid spurious frequencies associated with sharp discontinuities. All the vowels were normalised to the same RMS (0.1, relative to ±1 wav format amplitude) and to the same voice pitch of 113 Hz (corresponding to an average male). These five vowels comprise what is referred to as the 'canonical' vowels.

The ‘size’ aspect of vowel sounds is determined by Vocal Tract Length (VTL). The initial canonical vowels were manipulated to produce vowels with arbitrary VTL and pitch using the high-quality vocoder STRAIGHT (Kawahara, 1997). It performs pitch synchronous extraction of a complex version of the speech envelope which is independent of glottal pulse timing (pitch). This makes it possible to manipulate the VTL and re-synthesise with an arbitrary glottal-pulse rate. STRAIGHT mimics changes in VTL by taking the speech envelope and compressing/stretching it along the frequency axis. Therefore changes in VTL are affected within STRAIGHT by changes in the Spectral Envelope Ratio (SER). Small values of SER indicate lengthening of the vocal tract to simulate large adult males and large values of SER indicate shortening of the vocal tract to simulate children. The envelope codes information about the size of the VTL and it can be manipulated to produce the vowels of men, women and children from one initial vowel exemplar. The shape of the vocal tract, which determines the position of the formants in frequency, is heard as vowel type.

Following manipulation by STRAIGHT, the scaled vowels were subjected to further processing: the first 100 ms of the vowel waveform was removed because STRAIGHT takes some time to converge on the envelope; a cosine-squared gating function (10 ms onset, 30 ms offset, 465 ms plateau) was applied to the sounds, and the RMS was set to 0.025 (relative to ±1 wav format amplitude).

The software to control the experiments was written in-house using MATLAB 6.5 (Mathworks). Stimuli were played on a 24-bit sound card (Audigy 2, Sound Blaster) and then fed to a Tucker Davis Technologies (TDT 2) system running in passive mode. The stimuli were passed through an anti-aliasing filter with a sharp cutoff at 10 kHz. Stimuli were presented binaurally to the listener over a pair of AKG K240DF headphones. Listeners were seated in a double-walled, IAC sound-attenuating booth. The sound intensity of the vowels was 66 dB SPL.

B. Procedure and Listeners

Procedure: The vowel identification experiments were performed with a five-alternative, forced-choice paradigm in which the listener heard a scaled version of one of five stationary English vowels (/a/, /e/, /i/, /o/, /u/) and had to identify the vowel spoken by selecting the appropriate button on a response box displayed on a monitor in the booth. The vowel sounds were 500 ms in duration. No feedback as to whether the listener was right or wrong was given except at the beginning of the study. At the very start of the study we ensured that the listeners understood which button corresponded to which vowel sound, by playing 100 scaled vowels from within the range of everyday experience with feedback. The particular combinations of pitch and VTL in this set were not used in the vowel identification experiment.

The vowel identification data were gathered with two separate experimental paradigms whose names refer to the combinations of pitch and SER (mimicking VTL) of the stimuli: In the ‘*strip*’ paradigm (Fig. 3a), combinations of pitch and SER form a strip in the pitch/VTL space. There were eight strips presented in a different order to each listener to balance out any order/fatigue effects. Each strip consisted of nine (occasionally ten) combinations of pitch and SER for the five canonical vowels,

making a total of 9 (sample points) x 5 (vowels) x 10 (repetitions) = 450 trials per strip. The stimuli were presented in pseudo-random order in blocks of 45 trials (9 sample points x 5 vowels x 1 repetition). As a reminder of the canonical vowels (and to reinforce mapping of vowel heard to appropriate button), the set of five canonical vowels were presented with feedback at the start of the experimental run and thereafter every 100 trials. Each strip took approximately 30 minutes to complete. No feedback was given during data collection.

In the '*surface*' paradigm (Fig. 3b), there were ten runs, each consisting of all 49 combinations of 7 pitch and 7 SER values (making a total of 7 pitches x 7 SERs x 5 vowels = 245 trials per run). The pitches ranged over 6 octaves from 10 to 640 Hz. The SERs ranged over 2.5 octaves from 0.5 to 3.0[2]. Over the ten runs 2450 responses were collected per listener from which to generate a contour map of the vowel identification space. Listeners were reminded of the five canonical vowels at the start of the run and every 100 trials thereafter. Each surface took approximately 15 min to compete. No feedback was given during data collection.

The discrimination data were measured using two separate paradigms. One measured discrimination performance using single vowel instances and the other used a more complex speech-like stimulus task. These experimental approaches are referred to as the '*single vowel*' and '*speech-like*' discrimination tasks.

Single vowel discrimination functions were gathered for 9 points in the pitch/VTL space as shown in Fig. 9. The same vowels were used as in the vowel identification experiments. We used a temporal, two-alternative forced-choice paradigm with the method of constant stimuli. Each trial consisted of two intervals, one containing the standard stimulus and the other containing the test stimulus. The interval containing the standard stimulus was determined pseudo-randomly. The vowel type was the same across both intervals, i.e. /a/ was compared to /a/. The listener had to choose the interval containing the vowel spoken by the smaller speaker (discrimination of speaker size) or the interval containing the vowel spoken with the higher pitch (discrimination of voice pitch). No feedback was given. Six-point psychometric functions were measured with 10 trials per point (per listener). Each run consisted of five standard stimuli with their associated six test stimuli. Each experimental block consisted of thirty test stimuli and their associated standard stimuli presented in pseudo-random order. Each experimental run consisted of 300 trials (10 blocks of 30 test stimuli), taking approximately 30-40 min to complete. The listeners were given written instructions explaining the task in terms of pitch and speaker size. Most listeners considered it a natural task to judge the size of speaker of the vowel sounds. One listener thought it odd, but was able to do the task by thinking of the speaker as a cartoon character.

[2] An estimate of the size of speaker for a given SER can be derived by extrapolating from VTL versus height data in Fitch and Giedd (1999). An average adult male has a VTL of approximately 16 cm. An SER of 0.5 means that the spectrum envelope of the initial input vowel has been compressed by a factor of two while an SER of 3.0 means that the spectrum envelope has been dilated by a factor of three. Assuming linear scaling between formant position and VTL, our SERs are equivalent to VTLs of 32 cm (giants) and 5.3 cm (tiny children). Given the correlation between VTL and height (Fitch and Giedd, 1999 *cf*. Fig. 2a), our lowest SER of 0.5 would mimic the sound of a speaker the size of a giant 430 cm (14 feet) tall and our highest SER of 3.0 would mimic the acoustic properties of a speaker just 35 cm (1 foot 2 inches) tall.

In our fourth experiment, we again measured discrimination of speaker size but using a more speech-like task. As in our previous discrimination task, we used a temporal two-alternative forced-choice paradigm with the method of constant stimuli. However, in the new speaker size discrimination task each temporal interval consisted of a sequence of 4 of the 5 vowels (chosen randomly without replacement), following one of four pitch contours (rising, dropping, up-down, down-up), with different start pitches and where the intensity of all the vowels in one interval was roved in intensity (over a 6 dB range). The pitch values increased in quarter tone steps (~3% difference). We chose to randomly present 4 of the 5 vowels in each interval to discourage the listener from doing the discrimination by attending to one vowel only. The reason for using pitch contours was that we wished our vowel sequences to follow stereotypical pitch profiles of natural sentences, e.g. the rising pitch contour is meant to mimic the rising pitch of the interrogative sentence. By having different start pitches, we introduce more pitch variety to our task making it harder for listeners to track simple spectral cues across the two intervals. The reason for the intensity differences across intervals was to mirror speaker variability in voice intensity. Figure 4 shows a schematic of the experimental paradigm. The only consistent difference between the vowel sounds in the two intervals was the size of the speaker.

Listeners: Five listeners participated in each experiment. One listener was unable to take part in the discrimination experiments and was replaced by a new listener who took part in the first single vowel discrimination experiment only. Of the five listeners in our speech-like discrimination experiment, two had taken part in all the other experiments. The listeners ranged in age from 20 to 52 years. All had normal binaural hearing thresholds as tested by pure-tone audiometry using a two-alternative forced-choice adaptive procedure at frequencies of 0.5, 1, 2, 4 and 8 kHz.

III. RESULTS AND DISCUSSION

A. Vowel Identification Experiments

Strip Paradigm: The purpose of this experiment was to measure vowel identification performance in the pitch/VTL space both locally and densely. The other purpose of the experiment was to delimit the region of good performance of vowel identification which could then be measured using a regular lattice of sampling points in the subsequent surface vowel identification experiment. We sampled the pitch/VTL space along eight strips (Fig. 3a). Half of the eight strips were diagonal (strips 1-4) where the vowels were scaled in both pitch and VTL which will be referred to as Spectral Envelope Ratio (SER). The remaining half of the strips (strips 5-8) consisted of vowels scaled in either pitch or SER alone.

The results for the strip identification experiment are shown in Figs 5 and 6. Figure 5 shows the group psychometric functions for strip 1 (*cf*. Fig. 3a). Mean percent correct identification for each vowel is shown as a function of the pitch and SER of the vowel. The smooth curve through the data points is the best-fitting cumulative Gaussian (Foster and Bischof, 1997). In our 5AFC identification experiments, chance performance is 20% (d'=0.0) while identification threshold is 50% (d'=1.0). Performance is still at or above identification threshold to surprisingly low pitches (~25 Hz) and small SERs (0.52). Performance only drops to chance for very low

pitches (< 15 Hz) and very small SERs (< 0.45), i.e. as we move towards the extreme bottom left hand corner of the space (*cf*. Fig. 3a).

As the differences between vowels were generally small, the data were collapsed across vowels (as well as listeners) to show how performance falls off at the edges of the identification surface. The results for all 8 strips are presented in Fig. 6. In the condition where the glottal pulse rate was varied from 5 Hz to 20 Hz (*cf*. Fig. 3a, strip 5) while keeping the SER fixed at 1.0, identification performance was essentially perfect (Fig. 6(5)), despite the fact that all the stimuli were below the lower limit of melodic pitch of 33 Hz (Pressnitzer, Patterson and Krumbholz, 2001). For very high pitches (Fig. 6(7)), performance begins to break down where residue pitch fades away (Schouten, Ritsma and Cardozo, 1962). The three remaining diagonal strips (Fig. 6(2-4)) present differing patterns of results. Strip 4 (Fig. 6(4)) shows the rapid break down of performance as we move towards the bottom right hand corner (identification threshold at 235 Hz, 0.59 SER with chance performance at 280 Hz, 0.54 SER). Strip 3 (Fig. 6(3)) shows performance is remarkably resilient to scale vowel changes involving high pitches (up to 640 Hz) and high SERs (up to 2.83), with performance never falling below threshold. Strip 2 (Fig. 6(2)) stays above threshold until ~10 Hz and 2.44 SER, and then performance falls precipitously (chance performance at ~5 Hz and 2.88 SER). The two remaining strips show how performance changes as we hold pitch fixed and vary SER. Strip 6 (Fig. 6(6)) shows how performance breaks down at high SERs (implausibly small VTLs) and strip 8 (Fig. 6(8)) shows how performance breaks down at low SERs (implausibly large VTLs).

Surface Paradigm: The surface paradigm was intended to delimit the contours of vowel identification performance and in particular, the threshold contour which was the 50% correct identification contour (d'=1.0 in a 5AFC paradigm). Best-fitting cumulative Gaussians summarising performance in the strip experiments were used to choose the pitch-SER region over which to perform the surface experiment. The pitch-SER space was sampled regularly and reasonably densely as shown in Fig. 3b.

To aid comprehension, the results of the surface experiment are presented in two formats: Figure 7 shows vowel identification performance as a function of the log of pitch and the log of SER using a surface 2D plot in which grey tone shows mean percent correct. The sample points are shown as circles with interpolation between the data points. The heavy black line marks the 50% identification contour defined as threshold. For comparison, the range of pitch and SER in the normal population is shown by the ellipsoid superimposed on the 2D surface (Peterson and Barney, 1952; Fitch and Giedd, 1999; Huber *et al*, 1999). The data were averaged over the five subjects. Figure 8 presents the same data as a 3D wire-mesh surface with height showing mean percent correct. The wire-mesh is fixed (without interpolation) to the z-axis values of the 7x7 data points. The pitch-SER values sampled in the experiments are indicated by the circles on the 2D projection plane below the 3D wire-mesh surface. The 50% threshold is marked by the heavy black contour line.

Our results show that identification performance is above threshold for a wide range of pitch and SER values (Figs. 7 and 8). If good performance derived solely from experience, we might expect correct identification to assume an ellipsoid shape centred on the positive diagonal stretching from large adult male (pitch ~100 Hz, SER ~1) to small child (pitch ~265 Hz, SER ~1.4). Assmann *et al* (2002) have suggested

that vowel normalisation is based on learnt statistical correlation; that is, one learns that a child's /a/ is the same vowel token as an adult male's /a/ through experience, and despite the differences in pitch and formant frequencies. Our data show, however, that the normal range is a small sub-space of a much wider range where performance is well above threshold with essentially no training. This supra-threshold region in our data far exceeds the region of natural experience suggesting that performance is not reliant on experience but rather presents the operation of a normalisation process. If this process is a general property of auditory processing used to segregate structure and size information for *all* sounds (as suggested by Irino and Patterson, 2002), then it is not surprising that we can identify vowels with pitch and SER values beyond the normal range. In this view, vowel normalisation is a by-product of the auditory system's general scale-invariant properties.

The shape of the supra-threshold region is approximately rectangular (Fig. 7) which suggests that pitch and VTL are largely orthogonal in perception. It is generally assumed that pitch and VTL are decidedly correlated, with large adults having low 'deep' pitches and small children having high pitches. While the linkage between speaker *sex* and pitch is strong (adult males do tend as a group to have low pitches and small children do on average have high pitches), one cannot make reliable judgements about speaker size given only pitch. Thus pitch can be used to categorise sex but not to draw *intra*-sex inferences about speaker size (Bachorowski and Owren, 1999). This is understandable because unlike the vocal tract which is intimately related to the size of the cranium and hence body size, the vocal folds are not constrained by any body structure (Lass and Brown, 1978; Negus, 1949). We also routinely use prosody to make sentence distinctions, i.e. 'the baby is happy' with constant pitch is a statement while 'the baby is happy' with rising pitch is a question. Our results are consistent with a rough coupling of pitch and VTL (sex to size).

For SER values greater than 3 (implausibly small VTLs), performance breaks down for all pitches (10-640 Hz). Such a high SER value would place the third and higher formants beyond 5 kHz and would result in them being above the phase-locking limit. In a temporal model of perception where phase locking is important, it is not surprising that vowel identification begins to fail at such high SER values. In a spectral model of perception where phase locking is ignored, it is not clear why vowel identification should fail so long as the spectral pattern is within the range of human hearing.

B. Discrimination of Speaker Size and Voice Pitch

Having mapped vowel identification in the space of pitch and SER (VTL), we turned next to the question of the sensitivity to change along the two dimensions by measuring the just noticeable difference (JND) separately for voice pitch and speaker size (speaker size is the perceptual cue that listeners use to do the SER discrimination task). We measured the 'single vowel' speaker size JNDs at nine points and the voice pitch JNDs at five points in the pitch-SER space (Fig. 9). The points were chosen to determine whether discrimination performance was independent of vowel identification; that is, whether discrimination performance is only good when vowel identification is above threshold? More generally, does discrimination performance vary with vowel identification performance? Another discrimination experiment measured speaker size JND using a more speech-like stimulus sequence (*cf*. below).

Speaker size discrimination (single vowels): Figure 10 shows the group psychometric functions for single vowel speaker size discrimination in the middle of the pitch-SER space. Mean percent correct identification of the interval containing the smaller speaker is shown as a function of the SER of the test stimulus. The cue for the listener was the perceived size of the speaker. In this region (Fig. 9, point 5), listeners found this an entirely natural task. The point of subjective equality was typically within 1% of the physical standard value indicating accurate perception of speaker size (SER or VTL). The group psychometric functions for all vowels are monotonic and have relatively steep slopes indicating that performance is very similar for all of the listeners. Sensitivity to vowel SER change as measured by the JND[3] was 8.1% (std dev. ±1.0%) in the centre of the pitch-SER space. For comparison, the JNDs for sound intensity (loudness), light intensity (brightness) and chemical intensity (odour) are 5-10%, 14% and 25% respectively (Miller, 1947; Cornsweet and Pinsker, 1956; Gescheider, 1976).

The JND for single vowel speaker size is presented in Figure 11a for the nine points tested in the pitch-SER space. Where vowel identification is above threshold (*cf*. Fig. 9), speaker size discrimination is good (typically <10%). Where vowel identification performance is at or below threshold, speaker size discrimination performance worsens (*cf*. Fig. 9, points 3, 6 and 9). For instance, the JND for point 3 is a massive 52% compared to 8.1% for point 5. The fact that discrimination performance and identification performance are correlated suggests that listeners have to recognize an object before they can judge the size of the source.

Voice pitch discrimination: The JNDs for voice pitch discrimination at five points in the pitch-SER space (*cf*. Fig. 9) are presented in Figure 11b. The values were very small (<2%) when the pitch was greater than 100 Hz. The JND rises towards 10% when the voice pitch is 40 Hz; that is, as the pitch approaches the lower limit of musical pitch, which according to Pressnitzer *et al*. (2001) is about 32 Hz. The JND values are greater than those reported for pure tones in the range 250-4000 Hz. However, the JND rises as frequency decreases below 250 Hz (e.g. Sek and Moore, 1995).

Speaker size discrimination (speech-like vowels): We also measured the size of speaker JNDs for vowels presented in more speech-like sequences (compared to the single vowel presentation). The reason for using single vowel stimuli was to map discrimination performance in a pitch-SER space where we have already measured vowel identification performance. However, there are two drawbacks of using single vowel presentations. First, the ecological validity is weak. We do not usually determine a speaker's size by listening to them uttering isolated vowels at us. Second, we want to avoid allowing the listener the option of focussing on some simple spectral cue (say the frequency of the first two formants), and using that to order the stimuli in size. Though our listeners spontaneously perceived our stimuli as being spoken from

[3] The JND is defined as [(76%-50%)/50%]*100. The point of subjective equality (50% correct) represents the matching point where listeners cannot discriminate between stimuli because they are so similar (chance performance of d'=0.0 in our 2AFC experiment). The 76% correct point is the traditional discrimination criterion value (d'=1.0 in our 2AFC experiment) where listeners can reliably discriminate between stimuli. All values read off the best-fitting cumulative Gaussian (Foster and Bischof, 1997).

different sized speakers, we thought it desirable to generate sets of stimuli where simple spectral cue tracking would not be possible. The changes in pitch across the stimuli in our speech-like vowel sequences precluded the tracking of simple spectral cues.

Size of speaker JNDs were collected at enough points (17) within the pitch-SER space to allow us to profitably make a perceptual map of resolution. Figure 12 shows size of speaker JND as a function of the log of pitch and the log of SER using a 2D surface plot in which grey tone shows resolution. Small JNDs (better performance) are plotted in greys approaching white and large JNDs (worse performance) are plotted in greys towards black. The actual sample points are shown as circles with interpolation between the data points. The range of pitch and SER in the normal population is indicated by the ellipsoid superimposed upon the 2D surface (Peterson and Barney, 1952; Fitch and Giedd, 1999; Huber *et al*., 1999). The data are averaged over the five subjects.

From the perceptual map we can see that discrimination performance (resolution) is best at the centre of the normal speech range (JND of 6.6%). Discrimination performance worsens towards the corners of the space (bottom-left 17.5%, top-left 23.4%, top-right 31.1%). We were unable to measure discrimination performance in the bottom-right (640 Hz, 0.6740 SER), as the stimuli were too distorted. For the purposes of extrapolating towards this bottom-right corner we used the JND from the single-vowel discrimination. Typically, JND values for single-vowel discrimination were slightly better than for speech-like discrimination so our 2D perceptual map probably underestimates the deterioration of performance towards the bottom-right corner of the pitch-SER space. Discrimination performance follows a reasonably well-behaved pattern deteriorating from the centre outwards (Fig. 12). There is a large triangular region of pitches and SERs where discrimination performance is excellent (JNDs < 10%). This wedge of good performance is much greater than the range of pitch and SER encountered in normal speech as indicated by the ellipsoid. The 15% and 20% contour lines show that for a huge range of pitches and SERs discrimination performance is still relatively strong.

Comparison of the size of speaker discrimination map (Fig. 12) to the vowel identification map (Fig. 9), shows that discrimination performance declines with identification performance. This agrees with our single-vowel speaker size discrimination data (Fig. 11a) where performance also appeared to be linked with identification performance. This suggests that recognition of an object is necessary before inferences can be made about the object's size relative to the population. While both discrimination and identification maps show high levels of performance over a wide range of pitches and SERs, the fall-off in discrimination performance is tilted along the positive diagonal (Fig. 12) whereas the fall-off in identification performance is much more rectangular (Fig. 7).

IV. GENERAL DISCUSSION

We measured the ability of human listeners to identify vowels manipulated to simulate speakers with pitches and VTLs scaled way beyond the usual range in the population. We found that identification performance was above threshold (d'=1.0,

=50% in a 5AFC experiment) for a huge range of pitches and VTLs – an area approximately ten times greater than the usual range of variation in the population (Figs. 7 and 8). If performance was due to learnt statistical correlation, i.e. the learning of pitch and formant frequency associations in the normal range by a neural net (Assmann *et al*., 2002) then it might be expected that high levels of performance would be restricted to the training set range of pitches and VTLs. It seems unlikely that a neural net model would be able to generalise these learnt associations to recognize vowels over a range so much greater than the training set. However, if the auditory system performs some kind of active re-scaling to all input sounds at a relatively early point in the auditory system (Irino and Patterson, 2002) then we would expect humans to be able to recognize vowels across a range of pitches and VTLs far greater than that normally encountered in human speech. In this view, vowel normalisation would be a *by-product* of the auditory system's scale-invariant properties. Recently, it has been shown that an auditory model modified to include a Mellin transform stage is able to accurately classify vowels with a huge range of pitches and VTLs, in close agreement with our human vowel identification map (Turner *et al*., 2004). We believe that the ability to classify vowels in such a huge space when coupled with the close agreement between perceptual and modelling results (Turner *et al*., 2004), supports the hypothesis that the auditory system has an active normalisation mechanism and that this is the basis of vowel normalisation in humans.

Assmann *et al* (2002) also measured vowel identification performance for vowels manipulated in pitch and SER (mimicking VTL). They used a range of pitches and SERs of approximately 100-400 Hz and 1-2 respectively. This means they covered the usual range of variation as well as the supra-normal higher end (very small children). Assmann and Nearey (2003) have recently extended their initial observations by using both upward and downward shifts in spectrum envelope and pitch of the vowels of adult males and females and 7-year old children. They interpreted the fall in performance of their listeners outside the normal range of pitch and SER as evidence of learnt statistical correlation of vowel sounds. In our experiment we come to the opposite conclusion! To help reconcile this contradiction we note that Assmann *et al*. required their listeners to make a classification from 11 possible vowels (we only used 5 vowels). The performance of Assmann *et al*.'s (2002) listeners at worst was 40% correct classification which is equivalent to a d' of 1.24 in an 11AFC experiment. Therefore, the listeners in Assmann *et al*. (2002) were still performing above threshold according to our conservative threshold of d'=1.0. The performance of Assmann and Nearey's (2003) listeners dropped to below 32% (d'=1.0 in 11AFC) when the SER was 2 but only for children and adult female's vowels (men remaining at ~50%). Given that the SER value is *relative* to the VTL of the input vowel set then an SER of 2 relative to a child's VTL (as in Assmann and Nearey, 2003) would be approximately equivalent to an SER of 3 relative to an adult male's VTL (this study). At an SER of 3 our identification performance had also declined to chance (Fig. 7).

We measured the ability of human listeners to discriminate changes in the pitch of scaled vowels at a number of points in the pitch-SER space (Figs. 9 and 11b). As expected, the ability to discriminate pitch was very high (JNDs of 2% or better). At low pitches (40 Hz), the JNDs increase to about 10%. This means that generally we

are very sensitive to differences in voice pitch, a fact presumably linked to the use of prosodic variation to make various linguistic distinctions.

More interestingly, we measured sensitivity to changes in SER (perceived as changes in speaker size). As far as we know this is the first time the JND for speaker size perception has been measured. Fitch reports the consistent use of size in scaled vowels using a rating procedure (Fitch, 1994). We measured speaker size discrimination functions at a number of points in the pitch-VTL space, both for single vowels and for speech-like presentations (Figs. 9 and 12). We found that listeners hear speaker size in vowels scaled way beyond the region encountered in normal speech (Fig. 9). All listeners were sensitive to changes in SER (VTL) with JNDs typically below 10% (Fig.11a). However, the ability to judge speaker size deteriorates when the SER values are outside the region of good vowel identification (*cf*. Fig. 9).

Single-vowel presentation is important because it allows us to directly compare identification and discrimination performance for the *same* stimuli. However, the drawbacks are that discrimination of speaker size with single vowels lacks face validity and also allows the listener the option of tracking some simple spectral cue (such as the position in frequency of the first formant), to order the stimuli rather than directly accessing some perceptual dimension of size. To ameliorate these objections we used an approach where the vowels were presented in a more natural speech-like fashion (Fig. 4). By varying the pitch of the stimuli and which vowels were presented, we ruled out the use of simple spectral tracking cues. The size of speaker discrimination functions were calculated at enough points to make it worthwhile to create a 2D map of discrimination performance (Fig. 12). The well behaved deterioration of discrimination performance with deterioration in identification performance suggests that the information coding both aspects of performance is shared by a common neural representation. Furthermore, we suggest that listeners have to recognize an object before they can judge the size of its source. This seems entirely reasonable – before you can tell the size of some object you need first to assign it to some *class* of objects. We can then determine the object's size relative to the population of objects.

We can derive an estimate of the number of resolvable steps in speaker size by counting JNDs spanning the *entire* supra-threshold range (Fig. 7). Given the variation in speaker size JND (Fig. 12), we estimate there are approximately 15-20 JNDs in speaker size (e.g. 15-20 steps along the vertical dimension at 160 Hz). Larger speaker size JNDs at higher and lower pitches would mean less resolution along other vertical slices in the VTL dimension. Our data suggests about 8 JNDs would span the usual range of variation in speaker size in the population.

A. Future work

The spectrum envelope defining the VT transfer function is more defined for lower than for higher pitches. For instance, the child's transfer function is less clearly defined (reduced spacing of the harmonics of the fundamental frequency) than the adult male's (*cf*. Fig. 1). Thus low-pitched calls provide better resolution of the formants than high-pitched sounds (Ryalls and Lieberman, 1982). Indeed, the best 'filler' for accurately defining the VT transfer function should be noise or whispers

(Tartter and Bruan, 1994). Such spectral considerations would suggest that repeating our speaker size discrimination experiments with whispered vowels (where the larynx is held partially open allowing turbulent air flow from the lungs), would result in better discrimination JNDs because the information about the formants is increased. However, our temporal-based auditory model (e.g. Patterson, Allerhand and Giguère, 1995; Patterson, 2000; Turner *et al*, 2004 with added Mellin module) requires repeated presentations of the same sound to build up accurate stabilised auditory representations. We would thus predict that speaker size discrimination JNDs for whispered vowels would be *worse* rather than better compared to JNDs for sustained periodic vowels. We would also measure the existence region of reliable vowel identification for whispered vowels as we vary the SER. Generally, using whispered vowels allows us to decouple the correlation between VTL and pitch in order to determine the spectral contribution to vowel normalisation in the absence of temporal fine-structure components.

We have shown that we can identify stationary essentially periodic vowel sounds over a wide range of pitches and VTLs. This empirical finding is strengthened by modelling simulations (Irino and Patterson, 2002; Turner *et al*., 2004). We are also interested in whether performance can be extended to other phoneme categories: the diphthongs (which are non-stationary as the vowel changes over time); the sonorants (/r/, /l/, /m/, /n/, /y/, /w/); plosive consonants (/b/, /d/, /g/, /p/, /t/, /k/) and fricative consonants (/s/, /f/, /v/, /th/).

Finally, we are fascinated as to where in the brain the size-invariant transform may be instantiated. The obvious way to chase this question is to perform imaging experiments in both humans and non-humans with a range of stimuli (not just vowels or vocalisations). By varying the stimuli in size we hope to be able to coax the scale centre into revealing itself. We hypothesize that this centre will be early in the auditory system possibly in the Medial Geniculate Body. We are rigorously pursuing these projects with our collaborators.

V. CONCLUSIONS

In a series of psychophysical experiments we measured the effect of manipulating voice pitch and SER (mimicking VTL change) on vowel identification performance (Figs 7 and 8). We found that human listeners were able to identify scaled vowels over a huge range – an area some ten times greater than that encountered in normal speech. This is consistent with our hypothesis that the auditory system includes an active normalisation process capable of preserving shape information despite changes in scale (size). We believe a strong candidate for this scale transform would be the Mellin transform (Irino and Patterson, 2002).

We also measured sensitivity to changes in voice pitch and SER (perceived as speaker size). We found that listeners hear speaker size in scaled vowels well beyond the region encountered in normal speech. This is true both for single vowels (Fig. 11) and for sequences of vowels presented in a natural speech-like manner (Fig. 12). All listeners were sensitive to changes in SER (VTL), with JNDs typically below 10% for a wide region of pitches and SERs. The decline in speaker size discrimination appears to be linked to the decline in vowel identification performance (*cf*. Fig. 12 *vs*. Fig. 7).

ACKNOWLEDGEMENTS

This material is based upon work supported by EOARD under Contract No. SPC-024083. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of EOARD. Pilot work has been reported in Turner *et al.*, 2004. Some of the work has been reported in abstract form (Smith, Patterson and Jefferis, 2003).

REFERENCES

Assmann, P. F., Nearey, T. M., and Scott, J. M. (**2002**). "Modeling the perception of frequency-shifted vowels," Proceedings of the 7th International Conference on Spoken Language Perception, ICS.

Assmann, P. F., and Nearey, T. M. (**2003**). "Frequency shifts and vowel identification," Proceedings of the 15th International Congress of Phonetic Sciences. ICPhS.

Bachorowski, J., and Owren, M. J. (**1999**). "Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech," Journal Acoustical Society of America **106**, 1054-1063.

Beckford, N. S., Rood, S. R., and Schaid, D. (**1985**). "Androgen stimulation and laryngeal development," Ann. Otol. Rhinol. Laryngol. **94**, 634-640.

Cohen, L. (**1993**). "The scale transform," IEEE Trans. Acoust. Speech and Signal Process. **41**, 3275-3292.

Cornsweet, T. N., and Pinsker, H. M. (**1965**). "Luminance discrimination of brief flashes under various conditions of adaptation," Journal Physiology (London) **176**, 294-310.

Fairchild, L. (**1981**). "Mate selection and behavioural thermoregulation in Fowler's toads," Science **212**, 950-951.

Fitch, W. T. (**1994**). "Vocal tract length perception and the evolution of language," Unpublished Ph.D., Brown University.

Fitch, W. T. (**1997**). "Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques," Journal Acoustical Society of America **102**, 1213-1222.

Fitch, W. T. (**1999**). "Acoustic exaggeration of size in birds by tracheal elongation: comparative and theoretical analyses," Journal Zoology **248**, 31-49.

Fitch, W. T., and Giedd, J. (**1999**). "Morphology and development of the human vocal tract: A study using magnetic resonance imaging," Journal Acoustical Society of America **106**, 1511-1522.

Fitch, W. T., and Reby, D. (**2001**). "The descended larynx is not uniquely human," Proceedings of the Royal Society of London B **268**, 1669-1675.

Foster, D. H., and Bischof W. F. (**1997**). "Bootstrap estimates of the statistical accuracy of thresholds obtained from psychometric functions," Spatial Vision **11**, 135-139.

Fu, Q-J. and Shannon, R. V. (**1999**). "Recognition of spectrally degraded and frequency-shifted vowels in acoustic and electric hearing," Journal Acoustical Society of America **105**, 1889-1900.

Gescheider, G. A. (**1976**). *Psychophysics; Method and theory*. Hillsdale, N. J: Lawrence Erlbaum Associates.

Hast, M. (**1989**). "The larynx of roaring and non-roaring cats," Journal Anatomy **163**, 117-121.

Huber, J. E., Stathopoulos, E. T., Curione, G. M., Ash, T. A., and Johnson, K. (**1999**). "Formants of children, women and men: The effects of vocal intensity variation," Journal Acoustical Society of America **106**, 1532-1542.

Irino, T., and Patterson, R. D. (**2002**). "Segregating information about the size and shape of the vocal tract using a time-domain auditory model: The stabilised wavelet-Mellin transform," Speech Communication **36**, 181-203.

Kawahara, H. (**1997**). "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited," Proceedings IEEE International Conference on Acoustics, Speech & Signal Processing (ICASSP '97) **2**, 1303-1306.

Lass, N. J., and Brown, W. S. (**1978**). "Correlational study of speakers' heights, weights, body surface areas, and speaking fundamental frequencies," Journal Acoustical Society of America **63**, 1218-1220.

Ohala, J. (**1984**). "Vocal tract evolution and size exaggeration," Phonetica **41**, 1-16.

Miller, G. A. (**1947**). "Sensitivity to changes in the intensity of white noise and its relation to masking and loudness," Journal Acoustical Society of America **19**, 609-619.

Negus, V. E. (**1949**). *The Comparative Anatomy and Physiology of the Larynx* (Hafner, New York).

Patterson, R. D., Allerhand, M. H. and Giguère, C. (**1995**). "Time-domain modelling of peripheral auditory processing: A modular architecture and a software platform," Journal of the Acoustical Society of America, **98**, 1890-1894.

Patterson, R. D. (**2000**). "Auditory Images: How complex sounds are represented in the auditory system," Journal of the Acoustical Society of Japan (E), **21**, 183-190.

Peterson, G. E., and Barney, H. I. (**1952**). "Control methods used in the study of vowels," Journal Acoustical Society of America **24**, 75-184.

Pressnitzer, D., Patterson, R. D., and Krumbholz, K. (**2001**). "The lower limit of pitch," Journal Acoustical Society of America **109**, 2074-2084.

Riede, T., and Fitch, W. T. (**1999**). "Vocal tract length and acoustics of vocalization in the domestic dog, Canis familiris," Journal Experimental Biology **202**, 2859-2869.

Schouten, J. F., Ritsma, R. J., and Cardozo, B. L. (**1962**). "Pitch of the residue," Journal Acoustical Society of America **34**, 1418-1424.

Sek, A. and Moore, B. C. J. (**1995**). "Frequency discrimination as a function of frequency, measured in several ways," Journal Acoustical Society of America **97**, 2479-2486.

Smith. D. R. R., Patterson, R. D., and Jefferis, J. (**2003**). "The perception of scale in vowel sounds," British Society of Audiology, Nottingham **P35**.

Ryalls, J. H., and Lieberman, P. (**1982**). "Fundamental frequency and vowel perception," Journal Acoustical Society of America **72**, 1631-1634.

Tartter, V. C., and Bruan, D. (**1994**). "Hearing smiles and frowns in normal and whisper registers," Journal Acoustical Society of America **96**, 2101-2107.

Turner, R. E., Al-Hames, M. A., Smith, D. R. R., Kawahara, H., Irino, T., and Patterson, R. D. (**2004**). "Vowel normalisation: Time-domain processing of the internal dynamics of speech," in Dynamics of speech production and perception

FIGURES

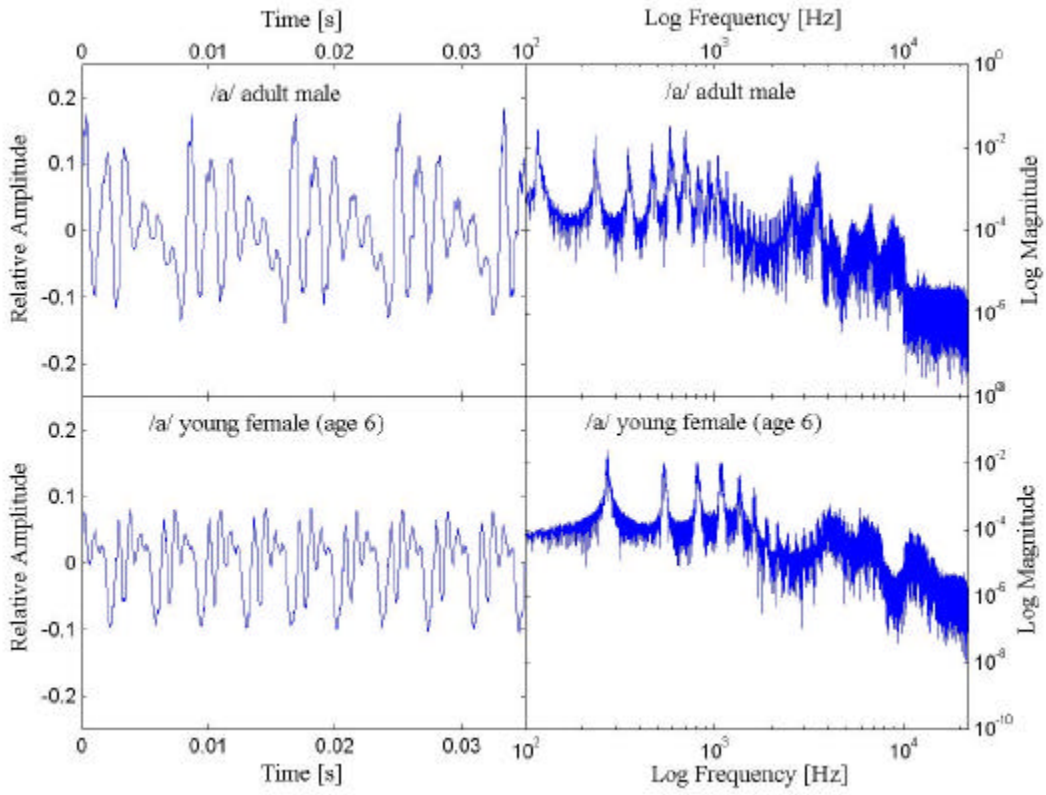

FIGURE 1. The same vowel /a/ as spoken by speakers of different size, sex and age. The top row shows the waveform and magnitude spectrum for /a/, as spoken by an adult male. The bottom row shows the waveform and magnitude spectrum for /a/, as spoken by a 6 year old female. The difference in the size of the larynx and length of the VT between the two speakers leads to two principal acoustical correlates. The greater mass of the vocal folds in the adult male mean that the fundamental frequency F0 (perceived as voice pitch) of the vocal fold oscillations is much lower than that of the female child, 117 Hz compared to 271 Hz respectively. The longer VT of the adult male means that the prominent frequencies (formants) are compressed in frequency compared to the formants of the young child. For instance, F1-F3 for the adult male are 619, 1018 and 2578 Hz compared to 937, 1336 and 3877 Hz for the young female.

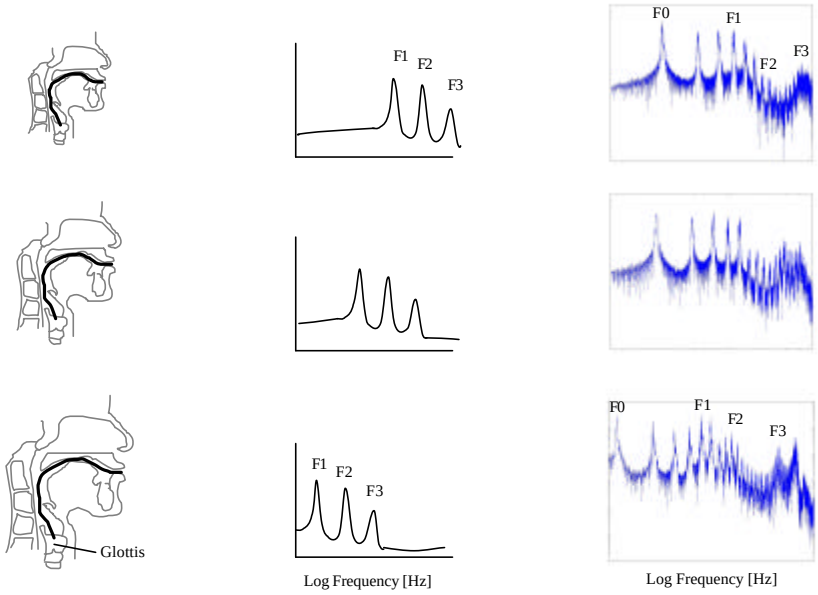

FIGURE 2. The affect of VTL upon the magnitude spectrum of a vowel. The first column shows cross-sections of the human vocal tract for child, adult female and adult male (first, second and third row). The vocal tract is shown by the bold line. The second column shows the magnitude spectrum of the filtering induced by the vocal tract with idealised formants F1-F3. The third column shows the result when the vocal tract is excited by a stream of glottal pulses.

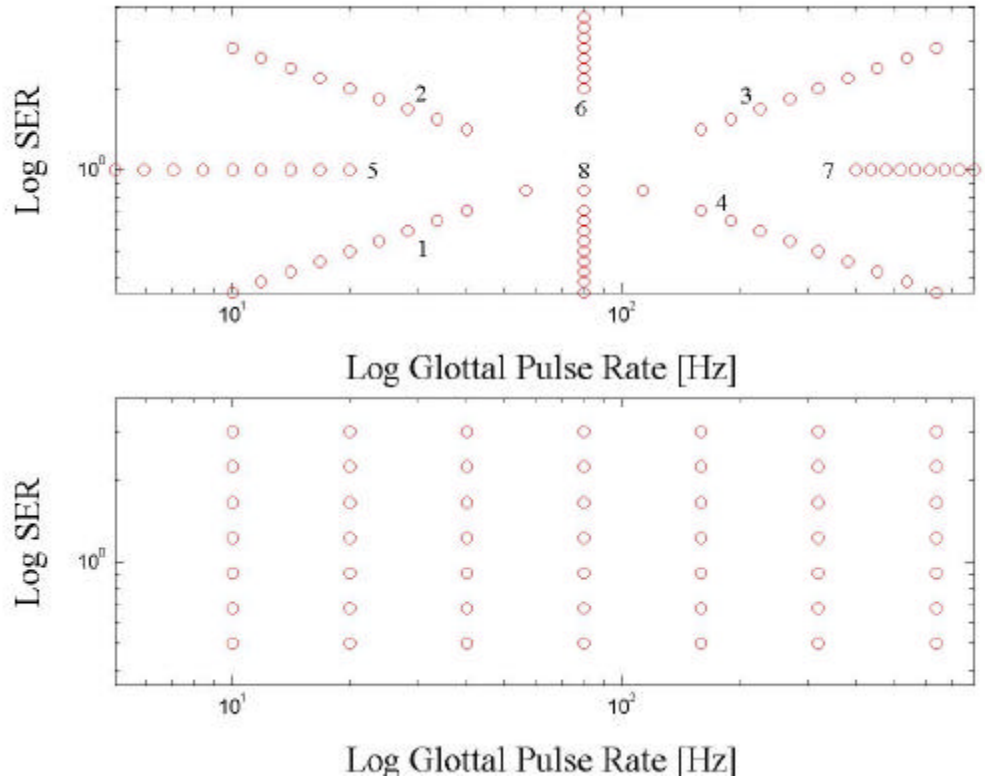

FIGURE 3. Experimental space of the vowel identification experiments. The horizontal axis is log glottal pulse rate in Hz. The vertical axis is log Spectral Envelope Ratio (SER). The interpretation of the horizontal axis is the vocal fold repetition rate (low for large adult males and high for small children). The physiological interpretation of the vertical axis is VTL. The SER determines the contraction or dilation of the spectral envelope applied by STRAIGHT during resynthesis (small SER values indicate lengthening of the VTL to simulate large adult males; large SER values indicate shortening of the VTL to simulate children). The top panel shows the eight *strips* of the first vowel identification experiment labelled 1-8. The circles show the particular combinations of pitch and SER used. The bottom panel shows the 7 x 7 sample points (10, 20, 40, 80, 160, 320 and 640 Hz by 0.5, 0.6740, 0.9086, 1.2247, 1.6510, 2.2255 and 3.0 SER) used in the second *surface* experiment.

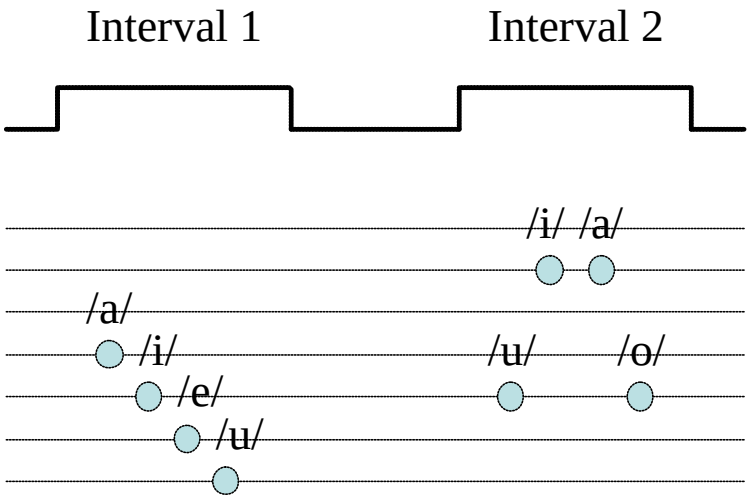

FIGURE 4. Speaker size discrimination task (speech-like). Each discrimination trial consisted of two temporal intervals where each interval was composed of a sequence of 4 of the 5 vowels (chosen randomly without replacement), following one of four pitch contours (rising, dropping, up-down, down-up), with different start pitches and where the intensity of all the vowels in one interval was roved in intensity over 6 dB range. Each pitch step differed by a quarter tone (~3%). The listener ($n=5$) had to chose the interval containing the vowels spoken by the smaller speaker. No feedback was given.

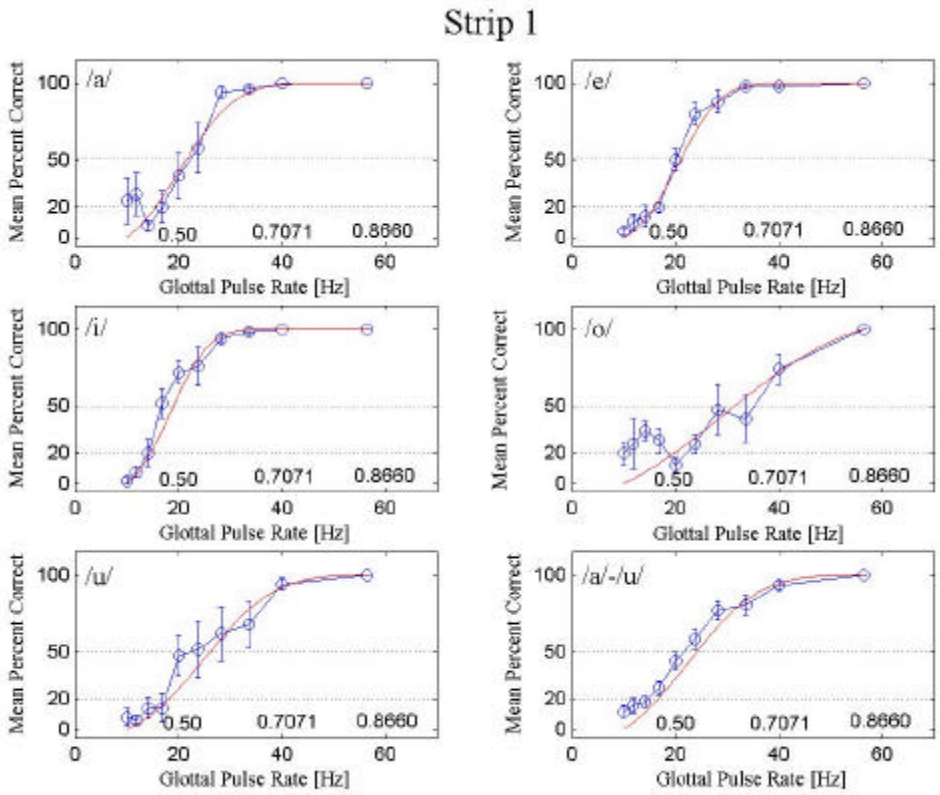

FIGURE 5. Vowel identification performance using the *strip* paradigm for strip 1 (*cf*. Fig. 3a). Mean per correct identification of the vowel is shown as a function of both the glottal pulse rate in Hz and Spectral Envelope Ratio (SER). The SER values are shown above the glottal pulse rate axis. Smooth curves through the data points are best-fitting cumulative Gaussians (Foster and Bischof, 1997). The horizontal dotted lines mark chance performance (20%, d'=0.0) and identification threshold (50%, d'=1.0) in our 5AFC task. The data are shown for each vowel separately. The means are based on the data of five listeners and each data point is based on 50 trials (10 trials from each listener). Error bars are standard error of the mean. The data, averaged across all five vowels, are shown on the bottom right (in this case each data point is based on 250 trials).

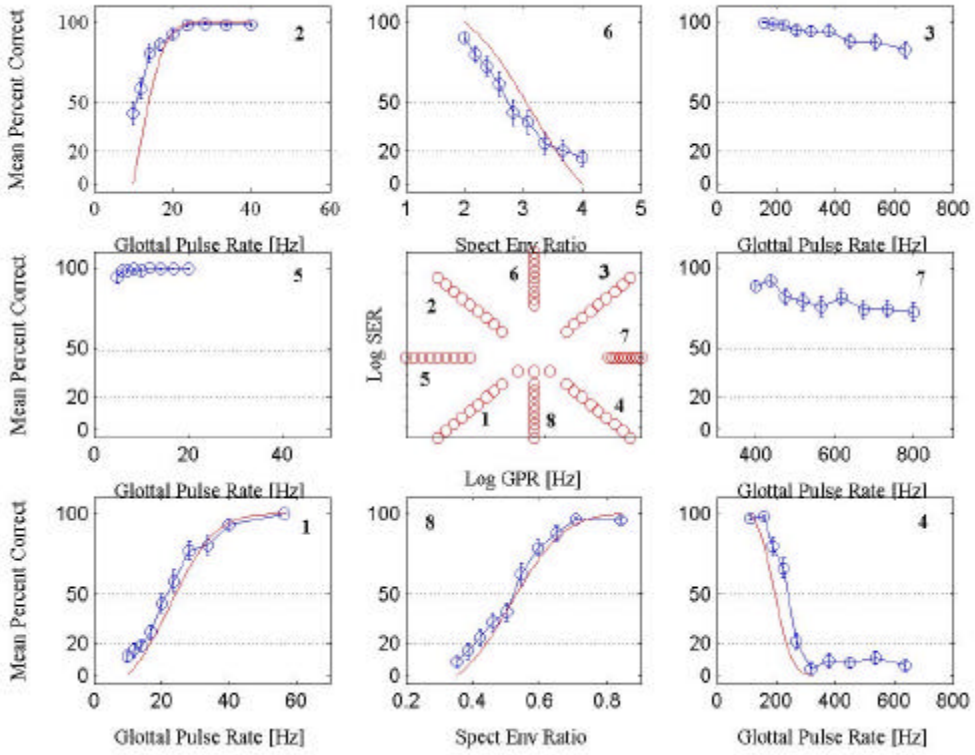

FIGURE 6. Vowel identification performance using the *strip* paradigm. Data collapsed across all five vowels and across all five listeners. Each data point based on 250 trials. Smooth curves are best-fitting cumulative Gaussians and have been used where appropriate. The centre panel shows the pitch-SER values for all eight strips. All other details as in Figure 5.

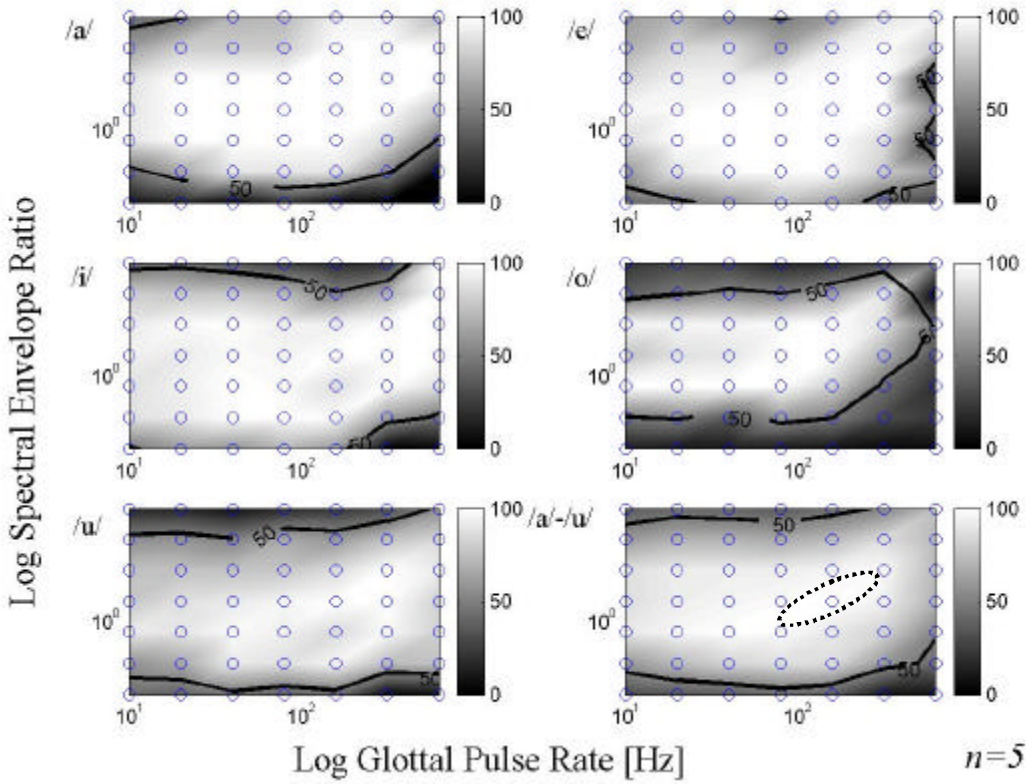

FIGURE 7. Vowel identification performance using the *surface* paradigm. The unit of the ordinate is the VTL of the original speaker (RP, male). The data are presented as a 2D surface plot with grey tone showing mean percent correct. Sample points are shown as circles, with cubic interpolation between data points. The means are based on the data of five listeners and each data point is based on 50 trials (10 trials from each listener). The data, averaged across all five vowels, are shown on the bottom right (in this case each data point is based on 250 trials). The thick black contour line marks the identification threshold (50%, d'=1.0) in our 5AFC experiment. The thin ellipsoid (bottom right panel) shows the range of pitch and SER values in the normal population (e.g. Peterson and Barney, 1952).

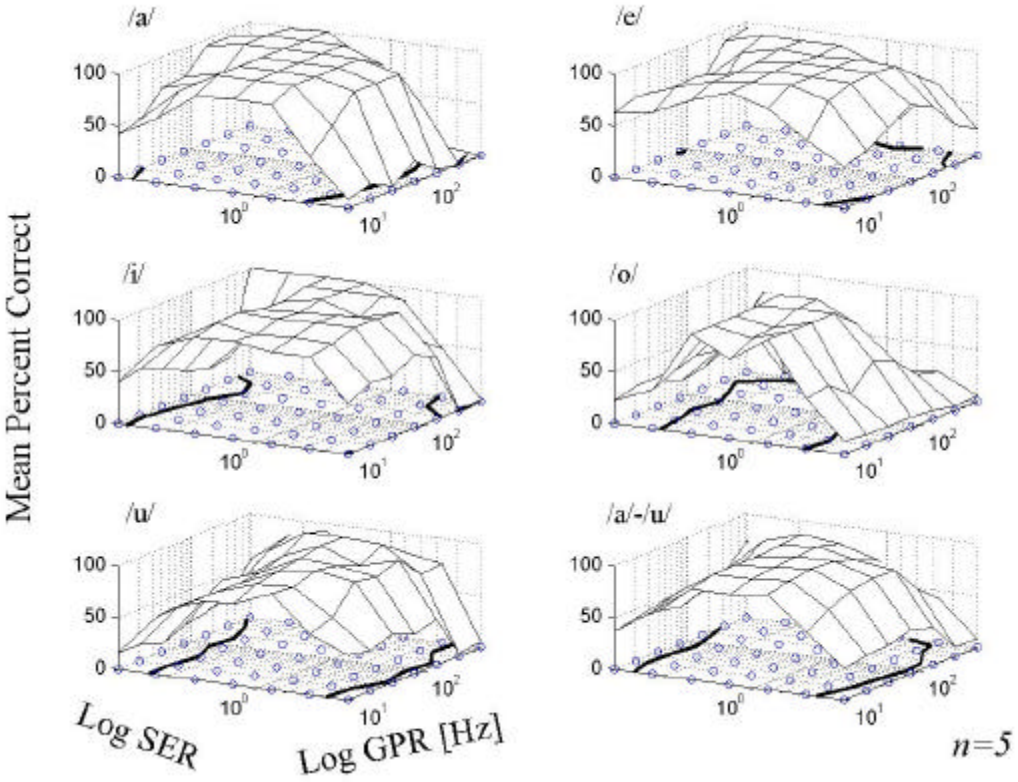

FIGURE 8. Vowel identification performance using the *surface* paradigm. In this figure, the data are presented as a 3D wire-mesh surface (no interpolation) with height showing mean percent correct. The glottal pulse rate and SER combinations used in the experiment are marked by the circles upon the 2D projection plane lying below the 3D surface. The identification threshold is mapped by the heavy black contour line on the 2D plane.

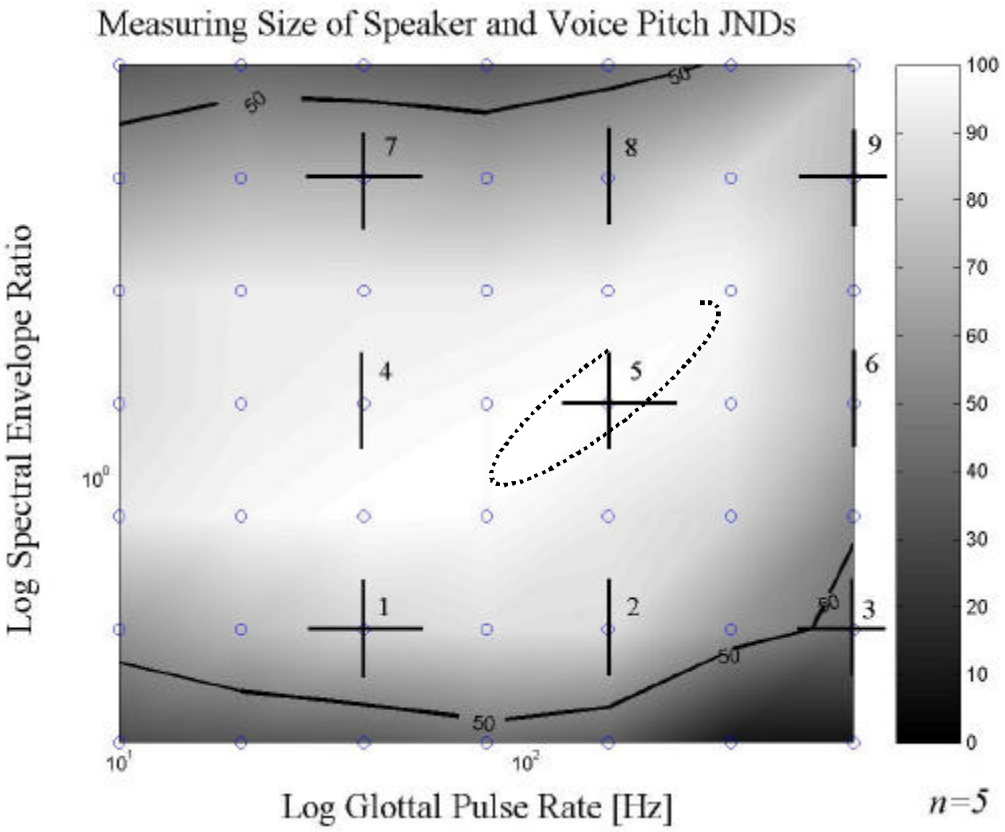

FIGURE 9. The points in the pitch-SER space at which speaker size and voice pitch discrimination performance were measured. The 2D surface plot shows vowel identification performance averaged across all five vowels and listeners (*cf*. Fig. 7, bottom right panel). The nine numbered points show the pitch-SER values of the standard stimuli used in the 2AFC experiments measuring the JND for speaker size and voice pitch. The ability to extract and use size of speaker information was measured at nine points in this space (1-9). The ability to extract and use voice pitch information was measured at five points in this space (1, 3, 5, 7, 9). The thin ellipsoid (bottom-right panel) shows the range of pitch and SER in the normal population (e.g. Peterson and Barney, 1952).

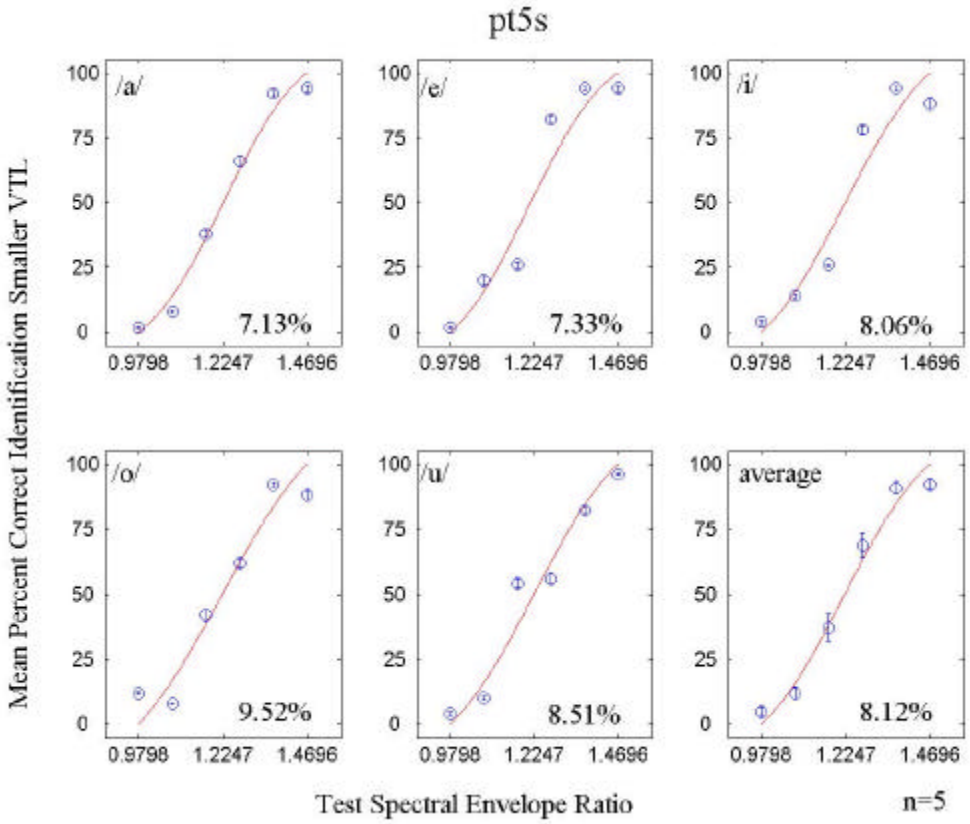

FIGURE 10. Speaker size discrimination in the centre of the normal speech range (*cf*. Fig. 9, point 5). Mean percent correct identification of the interval containing the smaller speaker (larger SER), as a function of test stimulus SER. Smooth curves through the data points are best-fitting cumulative Gaussians. The data are shown for each vowel separately, and averaged across all five vowels (bottom right panel). The means are based on the data of five listeners and each data point is based on 50 trials (10 trials from each listener). Error bars show the standard error of the mean. The data averaged across all five vowels are shown on the bottom right (in this case each data point is based on 250 trials). The JND calculated from the fitted curve is shown on the bottom right of each panel.

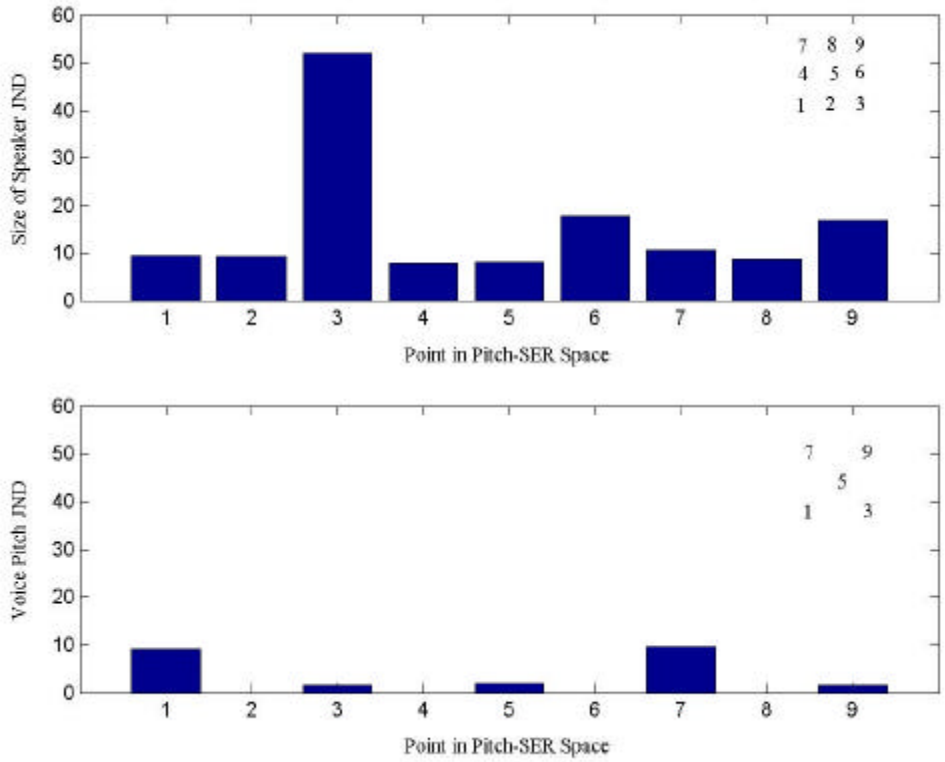

FIGURE 11. Size of speaker and voice pitch JNDs as measured at different points in the pitch-SER space (*cf.* Fig. 9). The position in the pitch-SER space is shown schematically in the top-right of each graph. Each JND value is based on a psychometric function fitted to 1500 trials (averaged across all five vowels and all five listeners).

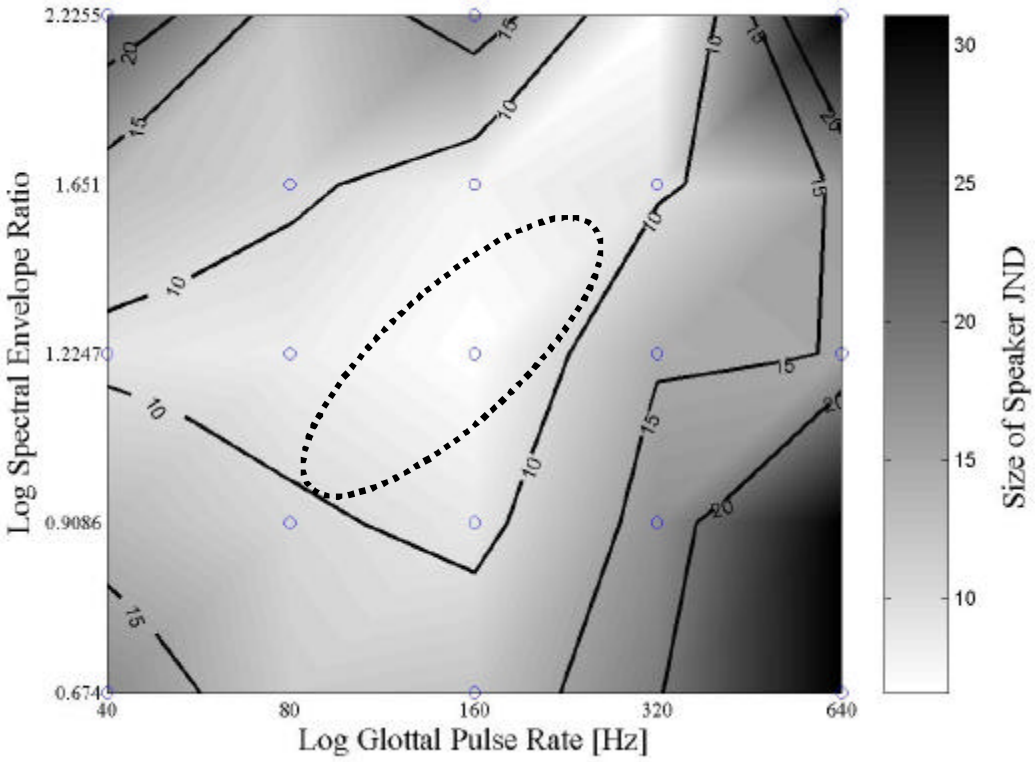

FIGURE 12. Map of size of speaker JNDs for speech-like sequences of vowels. The JNDs are presented as a 2D surface plot with grey tone showing discrimination performance. The JND was measured at the points shown by the circles, with interpolation between data points. The JNDs are based on a psychometric function fitted to 300 trials (averaged across all five listeners). The thick black contour lines mark the 10%, 15% and 20% JND discrimination thresholds. The dotted ellipsoid shows the range of pitch and SER values in the normal population (e.g. Peterson and Barney, 1952).